Object Positioning Based on Partial Preferences

Josep M. Mateo-Sanz[1], Josep Domingo-Ferrer[2], and Vicenç Torra

[1] Universitat Rovira i Virgili, Statistics Group,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
jmateo@etseq.urv.es

[2] Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
jdomingo@etse.urv.es

[3] Institut d'Investigació en Intel·ligència Artificial
Campus de Bellaterra, E-08193 Bellaterra, Catalonia
vtorra@iiia.csic.es

Abstract. In several situations, a set of objects must be positioned based on the preferences of a set of individuals. Sometimes, each individual can/does only include a limited subset of objects in his preferences (partial preferences). We present an approach whereby a matrix of distances between objects can be derived based on the partial preferences expressed by individuals on those objects. In this way, the similarities and differences between the various objects can subsequently be analyzed. A graphical representation of objects can also be obtained from the distance matrix using classical multivariate techniques such as hierarchical classification and multidimensional scaling.

Keywords: Preference structures, Object representation, Multivariate analysis, Classification.

1 Introduction

Preference structures have been an active area of research in the last years as they can be used to model preferences in a broad range of different applications. The appearance of the World Wide Web with a strong need for search engines and interactive tools for information access [9] has further magnified the importance of preference structures. Those structures are now pervasive in most systems to optimize access to knowledge.

The need for preference structures in real-world applications requires the development of new tools and methods. In this way, beyond basic research on preference modeling, new topics are of interest. For example:

- aggregation of preference structures to cope with metasearch engines (systems that search different databases);
- methods to compute similarities and distances between preferences (*e.g.* to cluster customers on the basis of their preferences);

V. Torra and Y. Narukawa (Eds.): MDAI 2004, LNAI 3131, pp. 252–259, 2004.

- methods to compute similarities and distances between alternatives considered in preferences (*e.g.* to cluster products on the basis of customers preferences).

Extensive research is documented in the literature on how to establish a sound basis for preference modeling. *E.g.* see [12] for a detailed description of results in this area; see [4] for a description of the field from a historical perspective; see also [11] for recent results in this area. Similarly, a large number of contributions have been devoted to preference aggregation, especially for retrieval from multiple sources or by multiple engines. See [9], [14] or [13] for details on such systems and on operators for aggregation of preferences. For a more technical paper on aggregation see *e.g.* [7].

One of the applications of modeling and aggregating preferences is to compute similarities and distances between objects from the preferences expressed on them. The availability of distances between objects allows positioning those objects and assessing their relationships. Computation of distances over preferences has been studied at length (see *e.g.* [10]).

1.1 Our Contribution

In several situations, a set of objects must be positioned based on the preferences of a set of individuals. Sometimes, each individual can/does only include a limited subset of objects in his preferences (partial preferences).

Example 1. Some real-life examples where partial preferences appear are the following:

- A set of individuals must choose their 15 favorite or most visited web pages among a large set of web pages or even all web pages.
- In order to analyze the consumer perception and preferences about several wine brands, a sample of consumers are asked to rank their favorite 5 wines.
- Students wishing to access higher education in a certain state are asked to rank their 8 favorite degrees among the degrees offered by the various universities in the state.

We present an approach whereby a matrix of distances between objects can be generated from the partial preferences expressed by individuals on those objects. In this way, the similarities and differences between the various objects can subsequently be analyzed. A graphical representation of objects can also be obtained from the distance matrix using classical multivariate techniques such as hierarchical classification and multidimensional scaling.

Section 2 contains basic concepts and notation used in the rest of the paper. The construction of the distance matrix between objects is specified in Section 3. A practical application is described in Section 4. Section 5 contains some conclusions.

2 Basic Concepts

The goal of this work is the representation of a set of objects from the preferences expressed by a set of individuals. The representation should be such that the similarities and the differences between objects become evident, *i.e.* that a relative positioning of objects arises.

Let us assume that we have a set of K objects and n individuals. Each individual chooses a subset of k objects among the K available objects and she ranks the k chosen objects according to his preferences. This yields an $n \times k$ preference matrix $\mathbf{X} = \{x_{ij}\}$, for $1 \leq i \leq n$ and $1 \leq j \leq k$, where x_{ij} represents the object ranked by the i-th individual in the j-th position in his order of preference.

Depending on the value of k, we have two types of preference matrices:

- *Total preference matrix.* If $k = K$, each individual expresses his preferences over the whole set of objects. Thus, every object appears in the preferences of every individual at some position.
- *Partial preference matrix.* If $k < K$, each individual only chooses a subset of k objects and ranks them according to his preferences. A specific object may not appear in the preferences of a specific individual.

We will concentrate here on partial preference matrices, so we assume $k < K$.

3 Construction of the Distance Matrix Between Objects

In order to represent the K objects, two classical techniques of multivariate analysis will be used [8]: hierarchical classification and multidimensional scaling. Those techniques require a matrix of distances between objects. We describe in this section the construction of a distance matrix from a matrix of partial preferences.

We start from the $n \times k$ preference matrix $\mathbf{X} = \{x_{ij}\}$. We go through the following steps to compute the distance between two objects r and s.

1. Compute a similarity measure between r and s which takes into account the difference between the positions of those objects in the order of preferences of the individuals. The closer the positions of objects r and s in the individuals' preferences, the more similar are the objects; conversely, the farther their positions, the less similar are objects. We present two possible approaches to implementing this similarity measure:
   (a) *Uniform distribution based on the distance between positions.* The idea is that the similarity between objects r and s contributed by an individual is proportional to the distance between the positions of preferences for those two objects expressed by the individual; if one or both objects were not chosen by that individual, then the similarity contribution for that individual is 0. Thus, the similarity sm^i_{rs} contributed by the i-th individual takes values in the interval (a, b), where $0 < a \ < b < 1$, if

both r and s were chosen by the i-th individual; it is 0 if r or s were not chosen by the i-th individual. One possible expression for sm_{rs}^i is as follows

$$sm_{rs}^i = \begin{cases} \frac{b(k-1)-a-(b-a)|j_r-j_s|}{k-2} & \text{if } r,s \in \{x_{i1},\cdots,x_{ik}\} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where j_r and j_s are the positions of objects r and s among the preferences of the i-th individual. Using Expression (1), if objects r and s are neighbors in the ranking (*i.e.*, $|j_r - j_s| = 1$), one obtains $sm_{rs}^i = b$ (high similarity); if objects r and s are at maximal distance (*i.e.*, $|j_r - j_s| = k - 1$), one obtains $sm_{rs}^i = a$ (low similarity).

(b) *Exponential distribution based on the distance between positions.* The idea is that the similarity between objects r and s contributed by an individual follows an exponential distribution between the positions of preferences for those two objects expressed by the individual. Thus, the similarity sm_{rs}^i contributed by the i-th individual can be expressed as

$$sm_{rs}^i = \begin{cases} \exp(-\alpha|j_r - j_s|) & \text{if } r,s \in \{x_{i1},\cdots,x_{ik}\} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where the parameter α is chosen to determine a specific exponential scale.

Regardless of whether uniform or exponential similarity is used, the overall similarity sm_{rs} between objects r and s is computed as $sm_{rs} = \sum_{i=1}^n sm_{rs}^i$.

2. Count the number q_{rs} of individuals who have chosen *both* r and s among their k preferences, regardless of their positions. Formally speaking,

$$q_{rs}^i = \begin{cases} 1 & \text{if } r,s \in \{x_{i1},\cdots,x_{ik}\} \\ 0 & \text{otherwise} \end{cases}$$

and $q_{rs} = \sum_{i=1}^n q_{rs}^i$. (Note that this computation does not make sense with a total preference matrix because one would have $q_{rs} = n$ for all r, s.)

3. Scale the similarity between objects r and s into the interval $[0, 1]$ as

$$s_{rs} = \left(\frac{sm_{rs}}{q_{rs}}\right)^{\frac{\bar{Q}}{q_{rs}}} \tag{3}$$

The rationale of Expression (3) is explained next:

- The ratio sm_{rs}/q_{rs} yields a similarity value scaled between 0 and 1.
- One would like to avoid high values of s_{rs} based on the choices of a very small number q_{rs} of individuals. That is the reason of the exponent in Expression (3): since $0 < sm_{rs}/q_{rs} < 1$, an exponent with a small q_{rs} reduces the value of s_{rs}, whereas an exponent with a large q_{rs} has an amplifying effect. The constant $\bar{Q}$ is the average of the nonzero values in the matrix $\mathbf{Q} = \{q_{ij}\}$, for $i, j = 1, \cdots, K$, and is used to scale q_{rs} in the exponent of Expression (3).

Table 1. Code, name, city and field of a subset of 23 university degrees offered in Catalonia

Code	Name of degree	City	Field
1	Biology	Barcelona	Science
2	Business administration	Barcelona	Social Sciences
3	Law	Barcelona	Social Sciences
4	Economy	Barcelona	Social Sciences
5	Humanities	Barcelona	Humanities
6	German translation and interpretation	Barcelona	Humanities
7	English translation and interpretation	Barcelona	Humanities
8	French translation and interpretation	Barcelona	Humanities
9	Audiovisual communication	Barcelona	Social Sciences
10	International trade	Barcelona	Social Sciences
11	Design	Barcelona	Engineering
12	Political and administration sciences	Barcelona	Social Sciences
13	Tourism	Mataró	Social Sciences
14	Industrial design	Barcelona	Engineering
15	Computer science	Barcelona	Engineering
16	Computer systems	Barcelona	Engineering
17	Business science	Barcelona	Social Sciences
18	Labor relations	Barcelona	Social Sciences
19	Business science	Mataró	Social Sciences
20	Architecture	Barcelona	Engineering
21	Business science (night)	Mataró	Social Sciences
22	Telematics	Barcelona	Engineering
23	Business science / Labor relations	Barcelona	Social Sciences

4. The distance matrix $\mathbf{D} = \{d_{ij}\}$, for $i, j = 1, \cdots, K$, can be derived from the similarity matrix $\mathbf{S} = \{s_{ij}\}$ computed in the previous step. There are several options for deriving distances from similarities [8]. For any two objects r and s, we list three possible derivations:

$$d_{rs} = 1 - s_{rs} \tag{4}$$

$$d_{rs} = \sqrt{1 - s_{rs}} \tag{5}$$

$$d_{rs} = \sqrt{1 - s_{rs}^2} \tag{6}$$

From the distance matrix $\mathbf{D}$ between objects, two classical techniques in multivariate analysis can be used:

1. *Hierarchical classification*([5, 6]). We choose the option of average linkage between groups to form new groups and obtain a dendrogram where the various objects are progressively clustered. The dendrogram is shaped like an inverted tree where leaves represent objects. The objects clustered at the lowest levels of the dendrogram (closest to leaves) are those with the highest similarity.

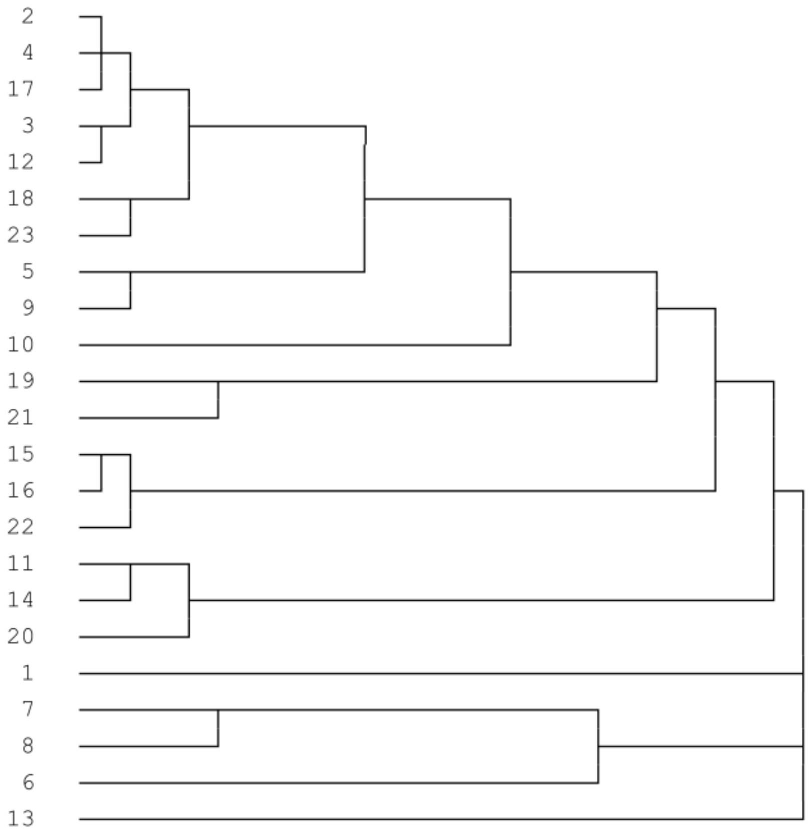

Fig. 1. Dendrogram of 23 Catalan university degrees

2. *Multidimensional scaling* ([1]). The interpretation of the relationships between objects is not obvious at all from direct examination of the distance matrix. The goal of multidimensional scaling is to plot in a graphic (typically in two dimensions) the structure of the distance matrix (or the similarity matrix). In this way, we obtain a simpler and clearer visualization of the connections between objects. In our empirical work, we have used the PROXSCAL algorithm [2, 3].

4 Application: Positioning University Degrees

Students wishing to enter the public university system of Catalonia are required to list their preferred degrees. Specifically, they can express up to $k = 8$ preferences, where each preference specifies a degree in a certain university; the student ranks his 8 preferences from most preferred (1) to least preferred (8). In the academic year 2003-2004, the 8 Catalan public universities offered $K = 378$

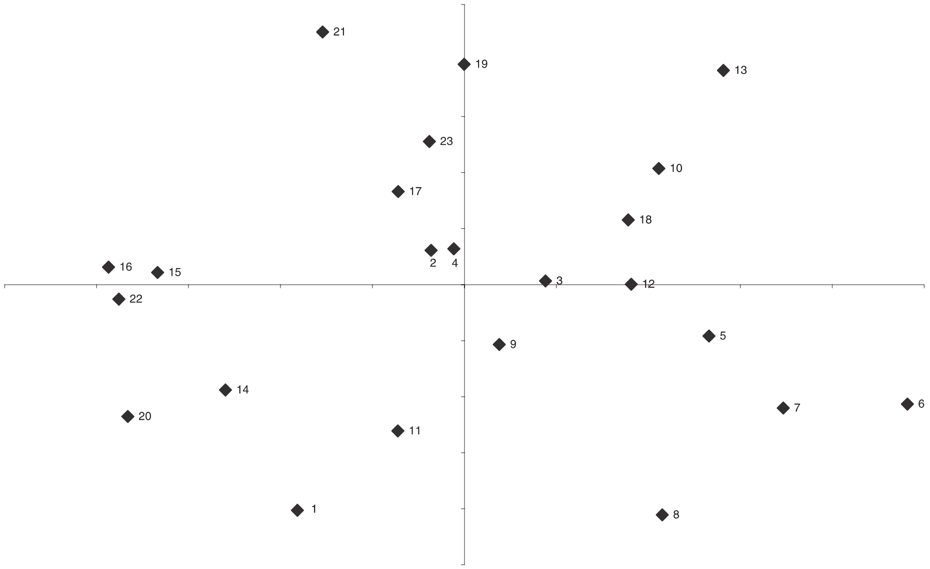

Fig. 2. Scatterplot of 23 Catalan university degrees

degrees. Also in that year, $n = 42125$ students entering the public university system expressed their (partial) preferences. Based on the full set of students' partial preferences, we have constructed the distance matrix $\mathbf{D}$ for the 378 degrees being offered.

In order to compute $\mathbf{D}$, uniform similarities have been used with an interval $(a, b) = (0.2, 0.8)$. The distance matrix has been derived from the similarity matrix by using the transformation described by Expresion (4).

The two aforementioned multivariate techniques (hierarchical classification and multidimensional scaling) have been applied to the distance matrix using version 11 of the SPSS statistical package.

For clarity and space limitation, we next present graphics corresponding to a subset of $k = 23$ degrees among the 378 total degrees. This subset of degrees are offered by Universitat Pompeu Fabra of Barcelona and are listed in Table 1. Figure 1 depicts the dendrogram obtained by hierarchical classification. Figure 2 depicts the two-dimensional scatterplot obtained by multidimensional scaling. Clearly, early clustered leaves in Figure 1 and/or close points in Figure 2 represent degrees perceived as similar. From the representations in the two figures, interesting inferences can be made: for example, two different degrees located at the same city are perceived as being closer than two equal university degrees located in distant cities. Such is the case for degrees 17 and 19, which are the same (business science) but located in two different cities; however, 17 is closer to different degrees in the same city, like 2 and 4. This might be seen as an indication of the scarce mobility of students.

5 Conclusion

Starting from a matrix of partial preferences expressed by a set of individuals on a set of objects, a matrix of distances between the objects has been constructed. The distance matrix already gives an idea about the relative positioning of the various objects from the point of view of the individuals. Thus, with this distance matrix, “close” and “far away” objects can readily be identified.

If multivariate techniques like hierarchical classification and multidimensional scaling are used on the distance matrix, a better visualization of the relative positioning of objects is obtained. Those techniques yield dendrograms and scatterplots which are in fact “maps” representing how the set of objects is perceived by the individuals.

References

[1] Borg, I., Groenen, P. W. (1997), *Modern Multidimensional Scaling: Theory and Applications*. Berlin: Springer. 257

[2] Busing, F., Commandeur, J., Heiser, W. J. (1997), PROXSCAL: A multidimensional scaling program for individual differences scaling with constraints. In W. Bandilla & F. Faulbaum (eds.), *Softstat'97: Advances in Statistical Software*, vol. 6, pp. 67-74. Stuttgart: Lucius & Lucius. 257

[3] Commandeur, J., Heiser, W. J., (1993), Mathematical derivation in the proximity scaling (PROXSCAL) of symmetric data matrices. Research Report RR-93-04, Department of Data Theory, Leiden University. 257

[4] De Baets, B., Fodor, J., (1997), Twenty years of fuzzy preference structures (1978-1997), *Belg. J. Oper. Res. Stat. Computat. Sci.*, 37, 61-82. 253

[5] Everitt, B. S. (1993), *Cluster Analysis*. London: Edward Arnold. 256

[6] Hartigan, J. A. (1975), *Clustering Algorithms*. New York: Wiley. 256

[7] Jacas, J., Recasens, J., (2003), Aggregation of T-transitive Relations, *Int. J. of Intel. Systems*, 18, 1193-1214. 253

[8] Legendre, P., Legendre, L., (1998), *Numerical Ecology (2nd English ed.)*. Amsterdam: Elsevier. 254, 256

[9] Pasi, G., (2003), Modeling User's Preferences in Systems for Information Access, *Int. J. of Intel. Systems*, 18, 793-808. 252, 253

[10] Spearman, C., (1906), Footrule for measuring correlation, *British Journal of Psychology*, 2, 89-108. 253

[11] Van de Walle, B., (2003), A Relational Analysis of Decision Makers' Preferences, *Int. J. of Intel. Systems*, 18, 175-791. 253

[12] Van de Walle, B., De Baets, B., Kerre, E. E., (1998), Characterizable fuzzy preference structures, *Ann. Oper. Res.*, 80, 105-136. 253

[13] Wu, Z., Meng, W., Yu, C., Liz, Z., (2001), Towards a higly scalable and effective metasearch engine, *Proc. 10th Int. Web Conf.*, Hong-Kong, 386-395. 253

[14] Yager, R. R., Rybalov, A., (1998), On the fusion of documents from multiple collection information retrieval systems, *J. Am. Soc. Inform. Sci.*, 49, 1177-1184. 253